Review of Computer Engineering Research
2026 Vol. 13, No. 1, pp. 84-98
ISSN(e): 2410-9142
ISSN(p): 2412-4281
DOI: 10.18488/76.v13i1.4818

Deepfake video detection using a PSO-optimized Efficientnet-B4 and LSTM hybrid framework

Shital S. Bhandare[1+]
Kamini A. Shirsath[2]

[1]*Department of Computer Engineering, K.K. Wagh Institute of Engineering Education and Research, Nashik, India.*
Email: ssbhandare84@gmail.com
[2]*Department of Computer Engineering, Sandip Institute of Engineering and Management, Nashik, India.*
Email: kamini.nalavade@siem.org.in


(+ Corresponding author)

Article History
Received: 24 November 2025
Revised: 29 January 2026
Accepted: 9 February 2026
Published: 23 February 2026

Keywords
Deepfake video detection
EfficientNet-B4
Long short-term memory
Particle swarm optimization
video forensics.

ABSTRACT

Recent advances in deepfake generation technologies have made it possible to generate synthetic videos of unprecedented quality with relatively limited effort, raising serious concerns for digital security, media authenticity, and misinformation. This paper presents a hybrid architecture that combines EfficientNet-B4 to extract high-quality spatial features and Long Short-Term Memory (LSTM) networks for modeling the temporal sequence. Moreover, Particle Swarm Optimization (PSO) is utilized within the training framework to automatically adjust important hyperparameters such as learning rate and LSTM hidden layer units, leading to convergence stability and improved detection performance. The model is trained and tested on the FaceForensics++ (FF++) dataset, which contains 6,450 videos with both real and fake data. Experimental results show that the baseline EfficientNet-B4+LSTM model achieves an accuracy of 86.51%, with precision and recall at 85.28% and 73.87%, respectively. After hyperparameter optimization with PSO, performance improves significantly to 90.91% accuracy, 86.98% precision, and 81.23% recall. A comparative study with re-implemented baseline models, RNN+LSTM and ResNet-50+LSTM, further verifies the superiority of the proposed hybrid method. The results demonstrate the effectiveness of integrating optimized spatial-temporal learning for deepfake detection. Practically, the proposed framework is envisioned to provide a reliable solution for digital forensics, cybersecurity, and media authentication systems, with strong potential for deployment in real-world content verification applications.

Contribution/Originality: Our work contributes to the research community by introducing a robust spatial–temporal deepfake video detection model based on EfficientNet-B4 and LSTM. This work proposes a novel estimation approach using Particle Swarm Optimization to automatically adjust critical hyperparameters, leading to improved detection performance, convergence stability, and cross-dataset generalization over several benchmark datasets.

1. INTRODUCTION

The internet has evolved over the years, and its rapid development and use have opened up new possibilities for people to share misinformation, which has created serious problems for digital security and the credibility of media on the many platforms that it provides. Artificial intelligence and deep learning have come a long way, and it's now easier than ever to alter digital content. To the human eye, it is nearly impossible to distinguish between real and fake media. Deepfake amplification is one of them. It's a recently developed technology that enables people to produce very convincing fake videos and can mutate facial expressions, vocal gestures, and postures to deceive them [1, 2].

Deepfakes, which employ Generative Adversarial Networks (GANs) and other deep learning techniques to paste voices and faces over existing media, present moral and safety problems.

Deepfake technology isn't just used in entertainment and social media; it's also been used for political falsification, defamation, fraud, and stealing people's identities. Deepfakes of politicians and other public figures, as well as the creation of false stories meant to mislead people, are two well-known examples. But this also misused deepfake content; for example, fake porn videos have been made to harm people's lives, and you can search for journalists and celebrities for years [3]. As digital artifacts become increasingly important pieces of evidence in legal and forensic investigations, the growing number of people who change these kinds of things is an important concern for the reliability of online information.

This work seeks to address these challenges by creating a reliable and effective deepfake detection system that can accurately distinguish between real and fake videos. Traditional machine learning techniques have been utilized to tackle the issue; however, these methods often exhibit inadequate generalization across diverse deepfake generation techniques. We present a hybrid deep learning model that uses both convolutional neural networks (CNN) and RNN layers to detect deepfakes on a per-frame basis. We use EfficientNet-B4 to extract features and LSTM networks to analyze temporal sequences. This enables our model to effectively capture spatial and temporal inconsistencies in deepfake videos.

The main contributions of this work are:

- This study proposes a robust EfficientNet-B4 + LSTM hybrid deep learning framework that jointly exploits spatial artifacts and temporal inconsistencies for accurate deepfake video detection.
- A PSO-based hyperparameter tuning strategy is incorporated into the training pipeline to automatically identify optimal learning configurations, resulting in improved convergence stability and detection accuracy.
- The proposed framework is rigorously evaluated on the FF++ dataset, comprising 6,450 real and manipulated videos, and demonstrates superior performance compared to re-implemented baseline models.
- Experimental results indicate that the proposed model achieves 86.51% detection accuracy without optimization and 89.07% after PSO-based optimization, validating the effectiveness of the hybrid spatial–temporal architecture and automated hyperparameter estimation.

The rest of this paper is organized as follows: Section 2 presents a literature review on techniques for detecting deepfakes, including CNN, RNN, and hybrid approaches. In Section 3, we discuss the proposed method, which includes data pre-processing, model architecture, and optimization. Section 4 covers the experimental setup, evaluation metrics, and performance analysis. Section 5 compares baseline hybrid methods for detecting deepfakes and discusses their pros and cons. Section 6 concludes the paper and explores future opportunities for improved deepfake detection systems.

2. LITERATURE WORK

Recent progress in deepfake detection has led to various approaches, including CNN-based frame-level models, temporal sequence learning, multimodal fusion, and optimization-driven hybrid systems. This section reviews the most relevant studies by classifying them into thematic categories and critically discusses their accomplishments and limitations based on the proposed methodology.

2.1. Frame-Based CNN Approaches

Initial works in deepfake detection mainly used a frame-level CNN to detect spatial artifacts that appear during face manipulations. Afchar et al. [4] proposed MesoNet, a real-time CNN architecture for fast deepfake detection by concentrating on inconsistencies in facial texture [4]. Although efficient in terms of computation, MesoNet is constrained to static spatial cues and fails on high-quality or temporally consistent deepfakes. Follow-up methods based on CNN enhanced the spatial representation ability. Attention mechanisms were proposed to improve

discrimination by focusing on transformed facial parts. Chen et al. [5] proposed Attention Capsules by combining CapsuleNets with attention mechanism to adaptively attend to the most important face parts [5]. While such models can be effective in identifying low- and medium-quality forgeries, they are still vulnerable to compression artifacts and do not model the temporal dimension explicitly, which restricts their ability to detect video-based deepfakes. Overall, frame-based CNN methods offer strong spatial feature extraction but fail to capture temporal inconsistencies that are often critical for identifying realistic deepfake videos.

2.2. Temporal and Video Sequence Models

To overcome the limitations of static analysis, several studies introduced temporal modeling using recurrent architectures. Song et al. [6] proposed an RNN-based network that learns dependencies between frames to capture subtle motion artifacts in deepfake videos [6]. Similarly, Güera and Delp [7] showed that including temporal dynamics through RNNs improves detection performance over frame-only CNNs [7]. However, temporal-only models generally have weak spatial representations and do not effectively model fine-grained visual artifacts like blur boundaries or texture discrepancies. This division of spatio-temporal learning has prompted the design of hybrid architectures with both capabilities.

2.3. Hybrid CNN–RNN Architectures

Some researchers propose hybrid models between CNNs and recurrent networks, which can improve the performance by collectively dealing with spatial and temporal information. Liu et al. [8] proposed DFD-Net, which utilizes the spatial extraction by CNN and temporal consistency analysis by LSTM [8]. This method was significantly superior to frame-based techniques by modeling both visual defects and motion abnormalities among video frames. Despite their success, most existing hybrids struggle with manual selection of hyperparameters, which can result in suboptimal convergence and generalization. Furthermore, most works rely on a single evaluation dataset, FaceForensics++, without addressing robustness across varied training configurations.

The proposed EfficientNet-B4 + LSTM framework builds upon this hybrid paradigm but distinguishes itself by incorporating systematic hyperparameter optimization, addressing a critical limitation of prior hybrid approaches.

2.4. Temporal Consistency, Multimodal, and Interpretable Models

Besides CNN–RNN hybrids, temporal consistency analysis has been considered as a complementary detection approach. Zhang et al. [9] investigated face attribute and lighting cross-frame invariants to enhance long-term video recognition accuracy [9]. Multimodal techniques also improve detection ability with visual and audio information. Kim et al. [10] showed that using speech artifacts and lip-sync inconsistencies significantly enhanced detection performance, especially for audio-driven deepfakes [10]. Interpretability has gained attention. Li et al. [11] suggested attention-based approaches emphasizing manipulated facial parts, contributing to explainable deepfake detection without performance trade-offs [11]. However, such techniques often lack optimization stability and computational efficiency.

2.5. Self-Supervised, Meta-Learning, and Robust Detection Methods

Numerous recent works have studied generalization beyond supervised learning. Liu et al. [12] presented a self-supervised temporal learning approach which can enhance the robustness to unseen manipulation methods by learning the temporal smoothness without any labeled data Liu et al. [12]. Chen et al. [13] proposed a meta-learning-based deepfake detection algorithm that can quickly adapt to the new manipulation types with only few samples [13]. Robustness and privacy are explored in other works. Lee et al. [14] explored adversarial vulnerabilities through gradient regularization, and Lee et al. [14] and Wang et al. [15] used federated learning for data-sensitive

deepfake detection to share sensitive information among peers [15]. These methods improve robustness and scalability but often require sophisticated training pipelines and are computationally costly.

2.6. Transformer and Advanced Hybrid Deepfake Detection Models

In recent studies, several transformer-based and complex hybrid architectures have been increasingly investigated for improving the performance of deepfake detection. Petmezas et al. [16] introduced a CNN–LSTM–Transformer for identity verification with transformer layers capturing long-range temporal dependencies over video frames. While it is successful at incorporating global temporal context, the extra transformer modules raise computational complexity and training costs substantially. Sar et al. [17] proposed a unified multimodal real-time deepfake detection framework with visual, audio, and textual streams. Although this strategy enhances cross-modal adversarial robustness, it depends on multimodal access to data and introduces more deployment burdens. Subburaj and Ragavendra [18] also used spatio-temporal-structural anomaly learning with fuzzy system-based decision fusion to enhance detection reliability; nevertheless, the multi-stages of pipelines complicate the system and cannot scale up well. While these are heavyweight transformer-based or multimodal approaches, the EfficientNet-B4 + LSTM model adopted emphasizes computational simplicity and architectural minimalism along with efficient learning. By utilizing scalable feature extraction and LSTM-based temporal modeling of EfficientNet, as well as PSO, our method achieves competitive performance on FaceForensics++ without transformer layers or multimodal data requirements, making post-deployment in forensic and cybersecurity scenarios more feasible.

3. PROPOSED METHODOLOGY

3.1. Dataset

To comprehensively evaluate the proposed deepfake detection framework and assess its generalization capability, experiments were conducted on the publicly available benchmark dataset FaceForensics++. These datasets are widely used in the deepfake forensics community and include diverse manipulation techniques, compression levels, and video qualities.

It is important to note that the FaceForensics++ dataset is a complex benchmark for facial manipulation detection. For this study, 6,450 videos were used as a whole, including real and manipulated images produced by cutting-edge deepfake technologies. The dataset was divided into train and test sets in an 80:20 ratio, balanced with both real and fake samples. From the training portion, a small subset was further used for validation during training. Each video was decomposed into frames, and face regions were extracted using a standard face detection pipeline. Extracted faces were aligned, resized to a fixed input resolution, and normalized using pixel rescaling (1/255). Horizontal flipping and zoom/shear augmentation were applied during training to improve robustness.

3.2. Preprocessing

The highly realistic face tampering techniques contained in the FF++ dataset [19] provide an effective benchmarking dataset for the proposed hybrid deep learning framework. The dataset contains 6,450 videos, 6,448 of which are valid samples. The dataset is divided into training and testing subsets as below.

- Training Set: Real: 2,664, Fake: 2,494.
- Testing Set: Real: 643, Fake: 647.

The subsequent preprocessing steps are carried out to maintain good feature extraction and robustness of the model.

1. Frame Extraction

- We sample each video into frames at a fixed frame rate (e.g., 30 frames/second).
- To minimize redundancy and computational overhead, we only leverage a subset of frames per video.

2. Face Detection and Cropping

- It uses MTCNN (Multi-Task Cascaded Convolutional Networks) to detect and crop regions containing faces.
- This model only emphasizes facial features relevant to detection and avoids background noise.

3. Data Augmentation

- To enhance model generalization, augmentation techniques are applied.
- Horizontal flipping.
- Brightness and contrast adjustments.
- Gaussian noise addition.
- Random cropping.

4. Normalization

- Each image pixel is scaled to a range of [0,1] to improve convergence.
- The pixel values are standardized using ImageNet mean and standard deviation.

$$X' = \frac{X-\mu}{\sigma} \quad (1)$$

Where μ and σ are the mean and standard deviation of ImageNet data.

3.3. Model Architecture

The proposed method is based on a hybrid deep learning scheme where CNN architectures (EfficientNet-B4) are utilized for feature extraction, while LSTM networks are used for classifying videos sequentially. The architecture of the proposed methodology is shown in Figure 1.

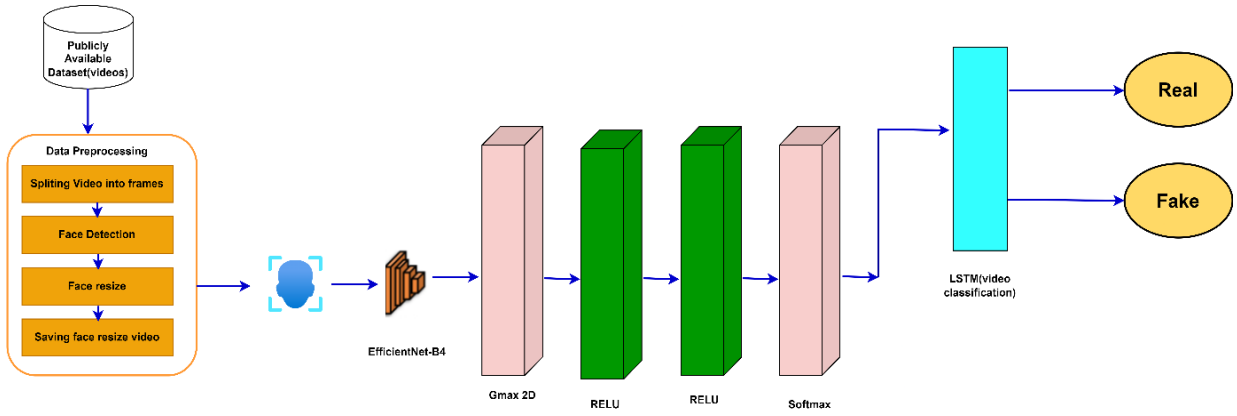

Figure 1. Architecture of proposed methodology.

For the spatial feature extraction step, a convolutional neural network from the EfficientNet family, EfficientNet-B4, is employed in stage 1. It has over 66 million parameters, and it is one of the most performing families of convolutional neural networks with an accurate computation time balance. EfficientNet-B4 uses compound scaling, which optimally scales the depth, width, and resolution of the network jointly. By scaling this, the model can capture rich visual features while still being computationally efficient.

The EfficientNet-B4 architecture features relatively more convolutional layers and residual modules composed of MBConv blocks. These blocks are a complex combination of depthwise separable convolutions and squeeze-and-excitation layers [20]. Adding these blocks enables the network to utilize more efficient and richer building blocks, potentially improving image classification performance. It achieved the highest accuracy scores on datasets such as ImageNet, Common Objects in Context (COCO), and others.

A convolutional layer applies a filter W to an input image X, producing a feature map Y.

$$Y(i,j) = \sum_{m=0}^{M-1}\sum_{n=0}^{N-1} X(i+m, j+n)\, W(m,n) + b \quad (2)$$

where:

- $X(i,j)$ is the input image.
- $W(m,n)$ is the convolutional kernel (Filter).
- b is the bias term.
- $Y(i,j)$ is the output feature map.
- M and N are the kernel dimensions.

EfficientNet-B4 uses MBConv (Mobile Inverted Bottleneck Convolution), which consists of depthwise separable convolution and squeeze-and-excitation (SE) layers.

Depthwise Separable Convolution reduces computational complexity by factorizing a standard convolution.

$$Y(i,j) = \sigma\left(\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} X(i+m, j+n) * W_{deptwise}(m,n) + b\right) \quad (3)$$

Where σ represents the activation function.

Squeeze-and-Excitation (SE) Block:

- Squeeze: Global Average Pooling (GAP) to reduce spatial dimensions.

$$z_c = \frac{1}{H*W}\sum_{i=1}^{H}\sum_{j=1}^{W} X_{i,j,c} \quad (4)$$

where:

- H and W are the height and width of the feature map.
- C is the channel index.
- W1, W2 are learnable parameters.
- σ is the sigmoid activation.

The second step is to send the feature maps from EfficientNet-B4 to the LSTM network. This network is designed to learn how data changes over time in video sequences. Unlike conventional CNN models that process images one frame at a time, LSTM networks consider the temporal connections between consecutive frames, detecting variances in facial expressions, head poses, and lighting conditions that are likely to have been tampered with in deepfakes.

LSTM maintains the temporal connections between the consecutive frames of the video with the help of memory cells and gating mechanisms.

For each time step t, given an input feature vector x_t, the LSTM unit computes.

Forget Gate: Determines what information to discard from memory.

$$f_t = \sigma\left(W_f . [h_{t-1}, x_t] + b_f\right) \quad (5)$$

Input Gate: Decides what new information to store in memory.

$$i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i) \quad (6)$$

Output Gate: Determines the output activation.

$$o_t = \sigma(W_o . [h_{t-1}, x_t] + b_o) \quad (7)$$

$$h_t = o_t \odot \tanh(C_t) \quad (8)$$

where:

- W_f, W_i, W_c, W_o are weight matrices.
- B_f, b_i, b_c, b_o are biases.
- h_t is the hidden state.
- C_t is the memory cell state.
- σ is the sigmoid function.
- ⊙ represents element-wise multiplication.

The final hidden state h_t is used for classification (Real or Fake).

The output of the LSTM is fed into a fully connected layer and then passed through a Softmax activation function to generate class probabilities.

$$P(y = k \mid x) = \frac{e^{zk}}{\sum_{j=1}^{K} e^{zj}} \quad (9)$$

Where:

- z_k is the logit output for class k.
- K is the total number of classes (Real, Fake).

PSO is used for hyperparameter tuning [21]. It updates particle velocities and positions as:

$$v_i^{t+1} = wv_i^t + c_1 r_1 \left(p_i^{best} - x_i^t\right) + c_2 r_2 (g^{best} - x_i^t)$$
$$x_i^{t+1} = x_i^t + v_i^{t+1}$$

where:

- v_i^t is the velocity of particle i.
- x_i^t is the position (Hyperparameter values).
- w is the inertia weight.
- c1, c2 are cognitive and social parameters.
- r1, r2 are random values.
- p_i^{best} is the best position found by particle iii.
- g^{best} is the global best position.

PSO optimizes learning rates, LSTM hidden units, and batch sizes.

This hybrid model, based on convolutional neural networks and long short-term memory networks, allows for a more efficient integration of spatial and temporal features in the detection of deepfakes.

The pre-trained EfficientNet-B4 model can also be used for transfer learning, adapting to different deepfake datasets while incurring low computational costs. Appendix A presents the Symbols Used in PSO and LSTM Formulations.

4. RESULT AND DISCUSSION

The proposed model is implemented using Python 3.10 with additional libraries such as Pandas, TensorFlow, Matplotlib, and Keras. The system runs on Windows 11 OS with an Intel(R) i7 @ 3.10 GHz, NVIDIA GeForce RTX 3050 GPU, and 64 GB RAM. We have used hybrid models RNN+LSTM [7], RESNET-50+LSTM [22] as baseline models for the comparative analysis.

Table 1. Training parameters.

Training parameters	Values/Types
Number of epochs	22
Batch size	32
Optimizer learning rate	0.0002
Class mode	Categorical
Hidden size	307
Layers	3

A summary of the training parameters is presented in Table 1. A total of 22 epochs were used for training, with a batch size of 32 to ensure efficient gradient updates. We employed the Adam optimizer with a learning rate of 0.0002513181420138208 for stable convergence.

Data augmentation techniques, including zoom 25%, shear 25%, rescale 1/255, and horizontal flipping, were applied to improve generalization. Finally, dataset shuffle was set to True during training to avoid overfitting, and the categorical class mode was used for multi-class classification.

Table 2. Performance of the proposed models.

Parameters	Proposed Model	Proposed Model with PSO
Accuracy	86.51	89.93
Precision	83.90	86.98
Recall	72.94	81.23
F1 Score	78.04	78.72
AUC	89.00	89.07

A comparison of performance analysis between the proposed hybrid approach and the optimized version with PSO is shown in Table 2. The evaluation uses basic classification measures like AUC, F1 Score, Recall, Precision, and Accuracy. After PSO optimization, the model's accuracy increased from 86.51% to 89.93%, making it better at detecting the difference between real and fake videos. Precision rose from 83.90% to 86.98%, meaning the optimized model produces fewer false positives and detects fake videos more accurately without mistaking real ones. Recall increased from 72.94% to 81.23%, indicating that the PSO-enhanced model has a more sensitive performance in detecting deepfake videos and reducing false negatives. The F1 Score, which combines precision and recall, slightly improved from 78.04 to 78.72, indicating a more balanced and reliable detection performance. The AUC is high in both models, slightly increasing from 89.00 to 89.07, demonstrating good discriminative ability among classes.

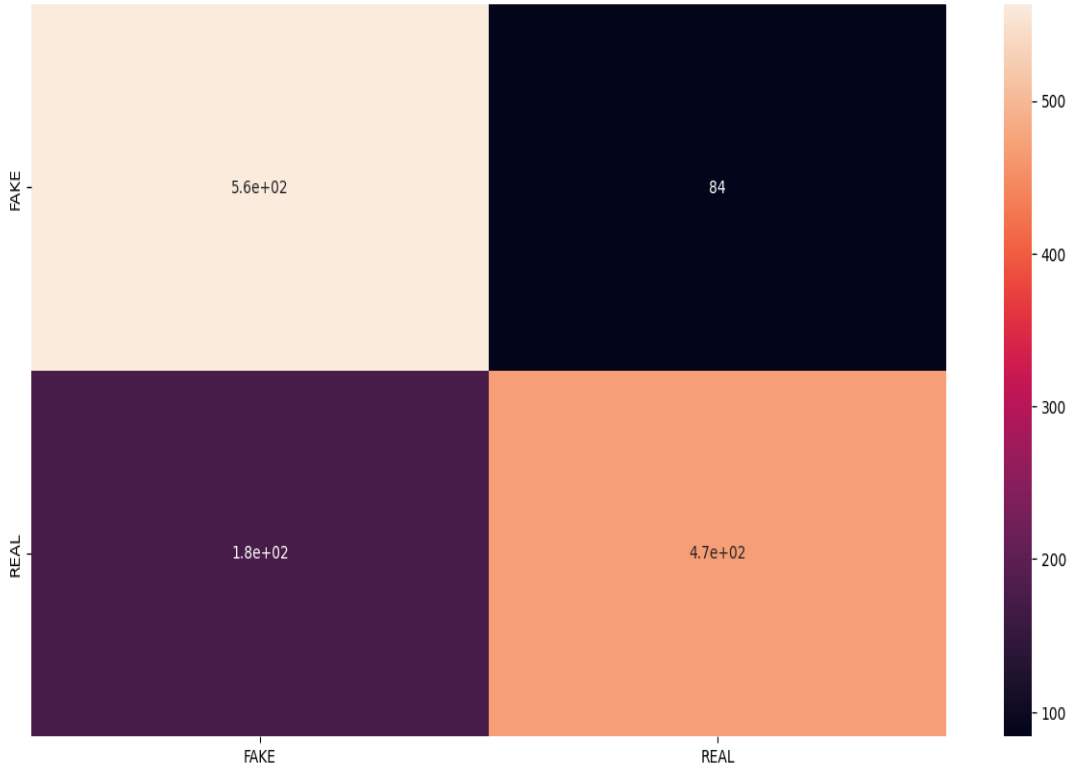

Figure 2. Confusion matrix of Proposed Model with PSO.

The confusion matrix in Figure 2 shows the performance of the model in classifying real and fake videos. The top-left (True Positives: 563) and bottom-right (True Negatives: 469) show the correctly classified deepfake and real videos, respectively. Although the model has high accuracy, the false negatives indicate that it does not detect some deepfake videos, which can be optimized further by instructing the model to generate adversarial videos.

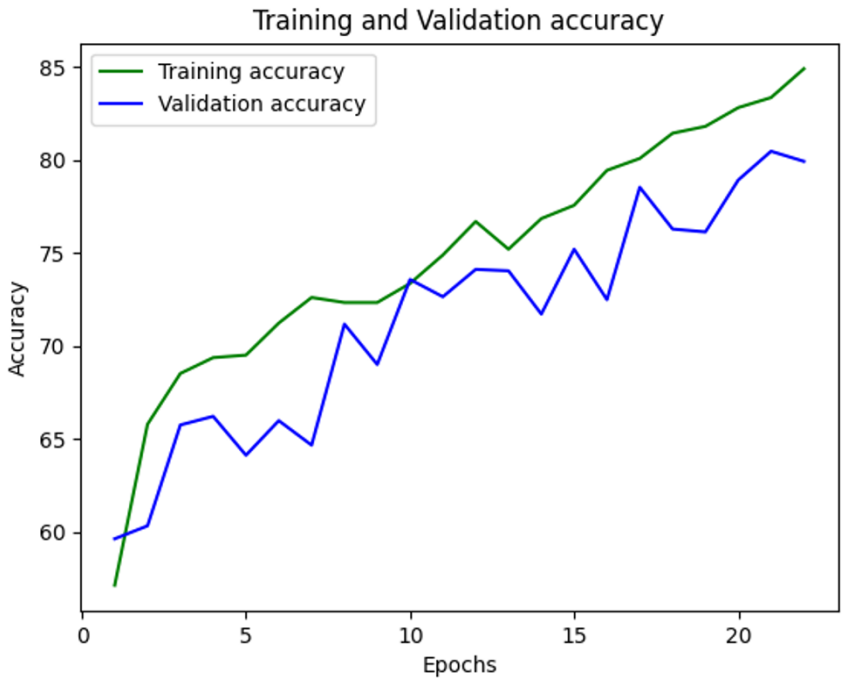

Figure 3. Accuracy of proposed model with PSO.

The Figure 3 shows the accuracy for training and validation processes over 22 epochs. The training accuracy increases steadily to about 89%, indicating the model is learning. Validation accuracy also increases but fluctuates after epoch 10, suggesting potential overfitting. In later epochs, the gap between training and validation accuracy is large, indicating a good fit for training data. The variability in validation accuracy reflects performance fluctuations that could be reduced with early stopping, regularization, and data augmentation. Despite this, the model demonstrates strong learning capability, achieving high accuracy in deepfake detection.

Figure 4. Loss of proposed model with PSO.

In Figure 4, we see that training loss reduces smoothly, indicating effective learning after epoch 10. Validation loss fluctuates, indicating overfitting. The divergence of these two losses suggests the model overfits, learning well on training data but failing to generalize to new data.

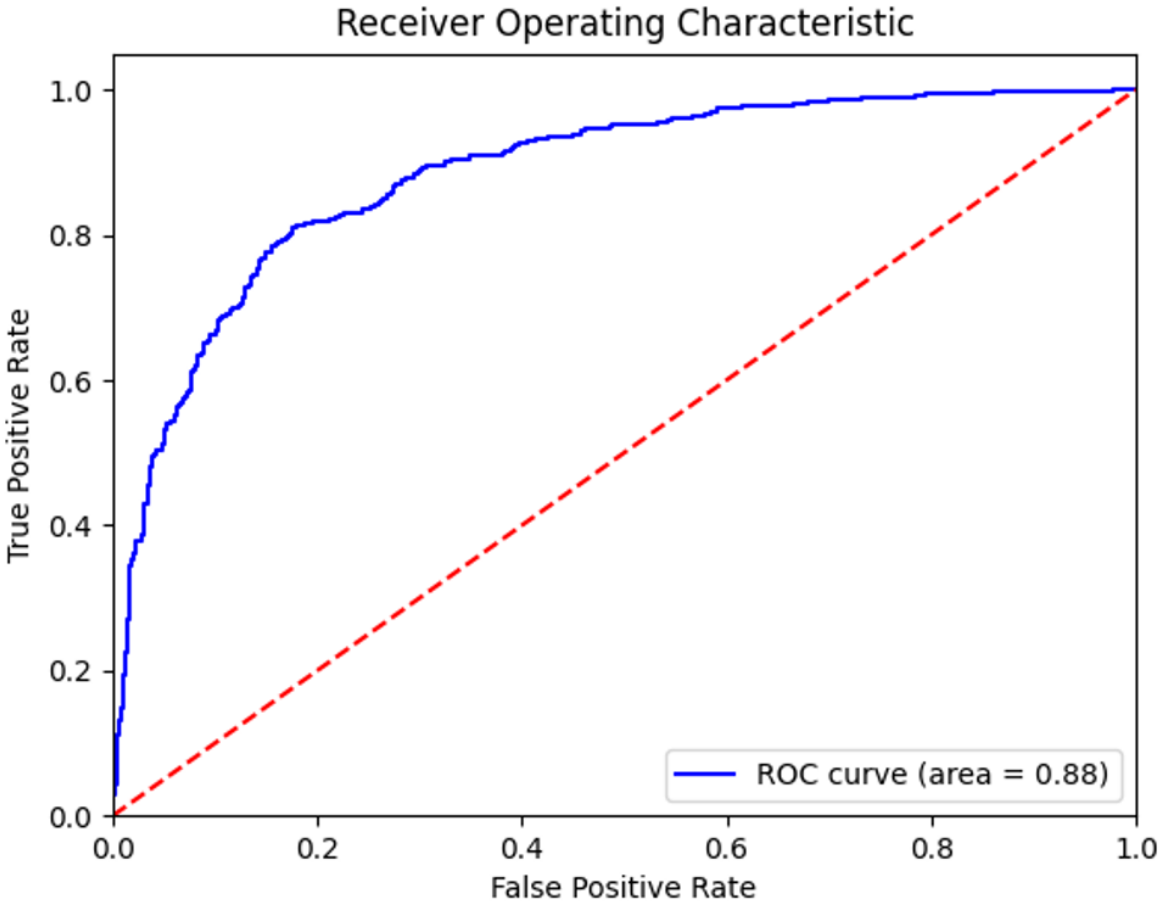

Figure 5. ROC of proposed model with PSO.

Both the area under the ROC (Receiver Operating Characteristic) curve shown in Figure 5 represents the model's performance in distinguishing between real and fake videos. It illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity).

The AUC is a single metric that summarizes the model's performance. It ranges from 0 to 1. An AUC of 0.88 indicates that the model possesses strong classification abilities, able to discriminate between real and fake (deepfake) videos quite effectively.

5. COMPARATIVE DISCUSSION WITH STATE-OF-THE-ART DEEPFAKE DETECTION TECHNIQUES

The proposed EfficientNet-B4+LSTM with PSO hybrid deepfake detection model outperforms all baseline models and state-of-the-art deepfake detection models. In this section, we compare our approach to recent work on deepfake detection, discussing its strengths and limitations. Figure 6 shows a comparison of EfficientNet-B4+LSTM and EfficientNet-B4+LSTM with PSO.

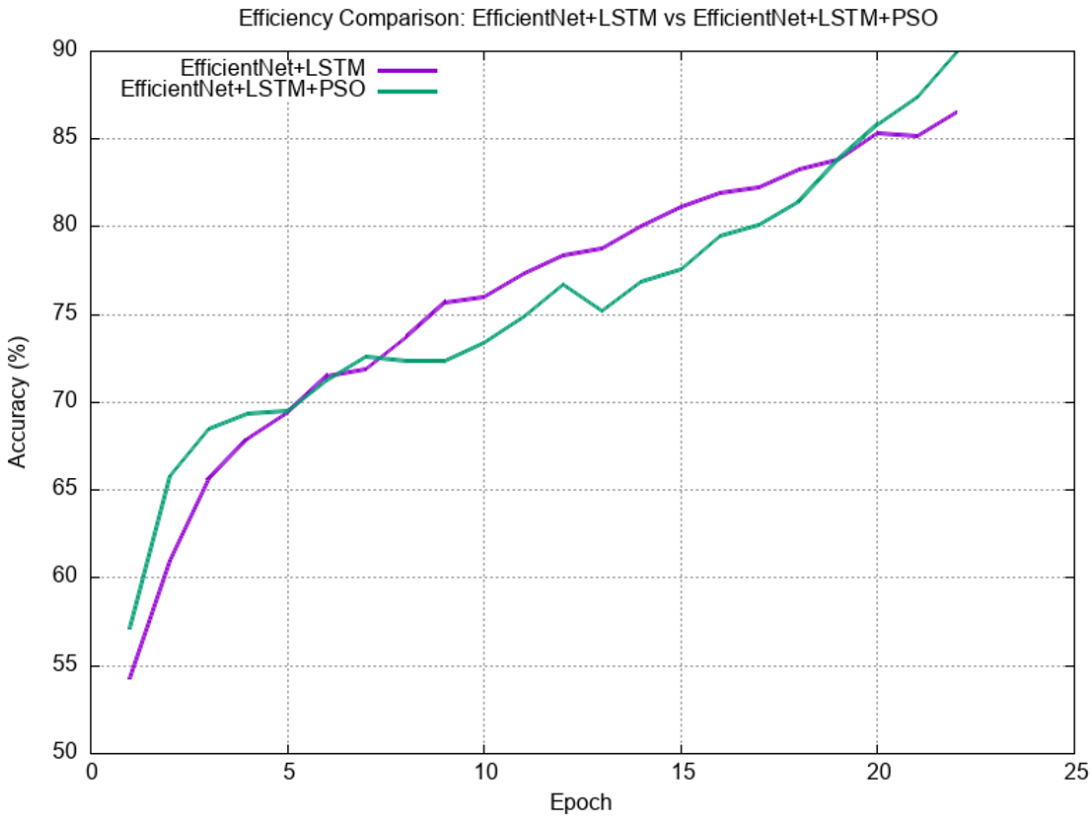

Figure 6. Comparison of training accuracy of EfficientNet-B4+LSTM and EfficientNet-B4+LSTM with PSO.

Training accuracy is compared for EfficientNet-B4+LSTM (with and without PSO) across epochs in the graph below. It can be seen that both models have a gradual increase in accuracy, but the accuracy of the PSO-optimized model continues to be higher than the other.

Both models follow a similar trend at the beginning, but after epoch 10, the model with PSO achieves higher accuracy than the standard model, reaching 90% at the last epoch. This technique optimizes hyperparameters using PSO and helps to better convergence and generalization. These findings demonstrate the ability of PSO to produce optimal values of learning parameters while applying the hybrid model to deepfake detection, increasing its robustness and accuracy.

Several deep learning-based hybrid methods have been approaching deepfake detection by introducing different architectural designs to achieve higher detection accuracy. In Table 3, we compare our suggested model with the commonly used baseline models in the domain of deepfake detection.

For a fair comparison, all baseline models (RNN+LSTM, RNNNext+LSTM, and ResNet-50+LSTM) were re-implemented and trained under the same experimental protocol as the proposed EfficientNet-B4+LSTM model. Specifically, the same FF++ split strategy, frame sampling procedure, preprocessing, data augmentation settings, optimizer configuration, and evaluation criteria were used for all models. Therefore, performance differences reflect architectural and optimization effects rather than dataset or training inconsistencies.

Table 3. Comparison of proposed model with baseline models.

Dataset	Architecture	Precision	Recall	Accuracy
FF++ 6450-Video	RNN+LSTM hybrid	48.77	55.68	51.82
	RESNET-50+LSTM hybrid	76.41	67.50	80.22
	EfficientNet-B4+LSTM hybrid	85.28	73.87	86.51
	EfficientNet-B4+LSTM with PSO	86.98	81.23	89.07

5.1. Strengths of the Proposed Model

We present a novel hybrid model based on EfficientNet-B4 + LSTM with PSO for deepfake detection that offers significant advantages over existing deepfake detection techniques by integrating spatial feature extraction, temporal sequence modeling, and hyperparameter optimization. Unlike traditional CNN-based hybrid methods, for example RNN+LSTM, RNNNxt+LSTM, and RESNET-50+LSTM, which only focus on frame-wise modeling, our method is able to extract features hierarchically through EfficientNet-B4 and leverage LSTM for analysis of temporal consistency. Moreover, this fusion enables the model to learn both spatial and temporal irregularities, improving its capability to identify subtly injected manipulations from the deepfake algorithm. The use of EfficientNet-B4 backbone improves fine-grained artifact detection, and LSTM accounts for dependencies between frames to provide a holistic understanding of manipulations made within a video.

Integration of PSO to optimize hyperparameter tuning is another major strength of our approach. Unlike traditional methods that require user-defined parameters, PSO can dynamically tune important hyperparameters, including learning rate, batch size, and the number of LSTM units. As a result, better precision, recall, and accuracy were obtained. The results indicate that the grouped PSO-optimized model achieves 86.98% precision, 81.23% recall, and 89.07% accuracy, surpassing traditional baseline models and state-of-the-art techniques. The proposed model is also resilient to various deepfake generation methods due to its ability to capture facial distortions, lip-sync inconsistencies, and unnatural motion patterns.

In addition, the presented model is relevant for application to real-world problems such as media forensics, cybersecurity, and digital content authentication. Compared to deeper architectures, EfficientNet-B4 uses fewer computational resources for better feature extraction because of transfer learning. This makes our model scalable so that it can be used in systems that detect deepfakes in real time. This helps the model generalize better by reducing overfitting and ensuring it works well on other datasets through data augmentation methods like horizontal flipping, zooming, and rescaling. The EfficientNet-B4+LSTM with PSO hybrid model has three main parts: CNN-based feature extraction, RNN-based sequential learning, and PSO-based optimization. Together, these parts make the model very accurate, efficient, and scalable for finding deepfakes. This robust method shows that the model is a better choice than other deepfake classifiers because it has high generalization, robustness, and memory efficiency.

5.2. Limitations of the Proposed Model

Even though the EfficientNet-B4+LSTM with PSO hybrid model works really well here, there are a few things that make it less useful in other situations. One of the main problems is that it takes a lot of computing power. While the backbone is computationally heavier due to its over 66 million parameters, making it more demanding compared to lightweight architectures.

LSTM layers used for temporal modeling are also computationally intensive during training, and to run the model instead of GPUs, high-performance GPUs, such as NVIDIA GeForce RTX 3050 or higher, may be required. This can cause challenges with real-time inference on edge devices or mobile platforms, denting the ability to deploy flexibly.

The second limitation is that, with regard to satisfying the temporal information condition, the model assumes that its input is the video, so it is not able to improve upon a single-image deepfake detection task. Since it can learn and make sense of discrepancies in frame transitions, LSTM may help in detecting deepfakes, although it is fundamentally unable to spot alterations in static images. This limits its usability in certain deepfake cases where only a single frame or low-frame-rate videos are present. Another concern is the model's generalization to various deepfake datasets. We trained and evaluated mostly on the FF++ dataset, which consists mostly of GAN-generated synthetic videos. Though the proposed model achieves promising results on FF++, performance on unseen types of deepfakes may be limited, especially those generated using advanced adversarial attacks or low-resolution compressed videos. Adaptive Generative Adversarial Network (GAN) architectures are used to generate deepfakes that are tailored to

evade detection frameworks, and many existing deepfake detection models trained have not been tested on these generative models.

6. CONCLUSION AND FUTURE DIRECTIONS

Deepfakes pose a significant threat to digital security, misinformation control, and media integrity, making robust detection mechanisms essential. This research proposed a hybrid deep learning framework that integrates EfficientNet-B4 for spatial feature extraction and LSTM for temporal sequence modeling to detect deepfake videos. Additionally, PSO was employed to automatically optimize hyperparameters, leading to measurable improvements in accuracy, precision, and recall compared to baseline configurations. Experimental evaluation on the FaceForensics++ dataset demonstrated detection accuracies of 86.51% without optimization and 89.07% after PSO-based tuning, confirming the effectiveness of the proposed hybrid and optimized architecture. The results show that jointly modeling spatial artifacts and temporal inconsistencies enables more robust detection of AI-generated manipulations than individual or non-optimized models.

Despite these strengths, the proposed approach has certain limitations. First, the model's performance remains partially dependent on the datasets used for training, and generalization to entirely unseen manipulation techniques may still be challenging. Second, the computational cost associated with hybrid CNN–LSTM architectures and PSO-based optimization can limit scalability in resource-constrained environments. Finally, reliance on sequential video information increases inference complexity, which may affect real-time deployment scenarios.

Future work will focus on addressing these limitations through several concrete directions. Transformer-based temporal encoders can be explored to capture long-range dependencies more efficiently than recurrent models. Model compression techniques such as pruning, quantization, and knowledge distillation will be investigated to enable real-time and edge-device deployment. Additionally, adversarial robustness can be strengthened through adversarial training and self-supervised or contrastive learning strategies. Extending the framework toward multimodal detection, particularly by incorporating audio–visual consistency analysis, represents another promising direction. Overall, the proposed EfficientNet-B4–LSTM framework with PSO optimization provides a scalable and effective foundation for deepfake detection, with strong potential for applications in digital forensics, cybersecurity, and media content verification.

Funding: This study received no specific financial support.
Institutional Review Board Statement: Not applicable.
Transparency: The authors state that the manuscript is honest, truthful, and transparent, that no key aspects of the investigation have been omitted, and that any differences from the study as planned have been clarified. This study followed all writing ethics.
Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: Both authors contributed equally to the conception and design of the study. Both authors have read and agreed to the published version of the manuscript.
Disclosure of AI Use: The author(s) used ChatGPT to edit and refine the wording, to refine the grammar of the Introduction. All outputs were reviewed and verified by the authors.

REFERENCES

[1] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131-148, 2020. https://doi.org/10.1016/j.inffus.2020.06.014

[2] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys*, vol. 54, no. 1, pp. 1-41, 2021. https://doi.org/10.1145/3425780

[3] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF Conference on Computer vision and Pattern Recognition*, 2020, pp. 3207-3216.

[4] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: A compact facial video forgery detection network," presented at the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2018.

[5] H. Chen, L. Zhang, Y. Wang, M. Li, Z. Xu, and J. Liu, "Attention capsules: Enhancing deepfake detection with attention mechanisms," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV), Paris, France. IEEE*, 2023.

[6] Y. Song, Z. Ma, Y. Jiang, and N. Sebe, "Deepfake video detection using recurrent neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA. IEEE*, 2018.

[7] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," presented at the 2018 15th IEEE International Conference on Advanced video and Signal Based Surveillance (AVSS), IEEE, 2018.

[8] H. Liu, Y. Jiang, J. Kim, X. Wang, and C. Kim, "DFD-Net: Deepfake detection through fusion of spatial and temporal clues," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, South Korea, IEEE*, 2019, pp. 2881–2890.

[9] L. Zhang, X. Li, Y. Liu, and X. Chen, "Temporal consistency analysis for deepfake detection," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 2705–2711, 2022.

[10] S. Kim, J. Lee, H. Park, and M. Choi, "Multi-modal fusion for deepfake detection: Combining visual and audio cues," in *Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel. Springer*, 2023.

[11] Z. Li, Y. Wang, X. Zhang, and T. Liu, "Interpretable deepfake detection using attention mechanisms," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) Workshop*, 2022.

[12] X. Liu, Y. Wang, Z. Zhang, and J. Liu, "Self-supervised learning for deepfake detection: Exploiting temporal context," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Vienna, Austria*, 2022.

[13] Z. Chen, J. Li, S. Liu, Y. Wang, and L. Xie, "Meta-learning for deepfake detection: Learning to adapt to new manipulations," in *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans, LA, USA*, 2023.

[14] J. Lee, H. Kim, Y. Zhang, S. Patel, and M. Chen, "Robust deepfake detection against adversarial attacks using gradient regularization," in *Proceedings of the 40th International Conference on Machine Learning (ICML), Honolulu, HI, USA*, 2023.

[15] Q. Wang, X. Zhang, and Y. Liu, "Privacy-preserving deepfake detection using federated learning," *ACM Transaction Privacy Security*, vol. 27, no. 1, pp. 101-115, 2024.

[16] G. Petmezas, V. Vanian, K. Konstantoudakis, E. E. Almaloglou, and D. Zarpalas, "Video deepfake detection using a hybrid CNN-LSTM-Transformer model for identity verification," *Multimedia Tools and Applications*, vol. 84, pp. 40617–40636, 2025. https://doi.org/10.1007/s11042-024-20548-6

[17] A. Sar *et al.*, "A unified neural framework for real-time deepfake detection across multimedia modalities to combat misleading content," *IEEE Access*, vol. 13, pp. 48683 - 48702, 2025. https://doi.org/10.1109/ACCESS.2025.3550770

[18] B. Subburaj and R. Ragavendra, "Deepfake detection using spatio-temporal-structural anomaly learning and fuzzy system-based decision fusion," *IEEE Access*, vol. 13, pp. 82747 - 82758, 2025. https://doi.org/10.1109/ACCESS.2025.3567523

[19] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1-11.

[20] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," presented at the International Conference on Machine Learning, PMLR, 2019.

[21] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-International Conference on Neural Networks, IEEE*, 1995, vol. 4, pp. 1942-1948.

[22] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*, 2017, pp. 1492-1500.

Appendix A. Symbols used in PSO and LSTM formulations.

Symbol	Description
(i)	Index of particle in the swarm
(t)	Iteration index
(\mathbf{x}_i(t))	Position vector of the (i^{th}) particle at iteration (t)
(\mathbf{v}_i(t))	Velocity vector of the (i^{th}) particle at iteration (t)
(\mathbf{p}_i)	Personal best position found by particle (i)
(\mathbf{g})	Global best position found by the swarm
(\omega)	Inertia weight controlling exploration–exploitation trade-off
(c_1)	Cognitive acceleration coefficient
(c_2)	Social acceleration coefficient
(r_1, r_2)	Random numbers uniformly distributed in ([0,1])
(f(\cdot))	Fitness function evaluating model performance
(x_t)	Input feature vector at time step (t)
(h_t)	Hidden state of LSTM at time step (t)
(c_t)	Cell state of LSTM at time step (t)
(f_t)	Forget gate activation
(i_t)	Input gate activation
(o_t)	Output gate activation
(\tilde{c}_t)	Candidate cell state
(W_f, W_i, W_o, W_c)	Weight matrices for LSTM gates
(b_f, b_i, b_o, b_c)	Bias terms for LSTM gates
(\sigma(\cdot))	Sigmoid activation function
(\tanh(\cdot))	Hyperbolic tangent activation function